W382–W384 *Nucleic Acids Research, 2006, Vol. 34, Web Server issue*
doi:10.1093/nar/gkl272

SWAKK: a web server for detecting positive selection in proteins using a sliding window substitution rate analysis

Han Liang, Weihua Zhou[1] and Laura F. Landweber[2,*]

Department of Chemistry, [1]Princeton Plasma Physics Laboratory and [2]Department of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ 08544, USA

Received February 14, 2006; Revised March 19, 2006; Accepted April 3, 2006

ABSTRACT

We present a bioinformatic web server (SWAKK) for detecting amino acid sites or regions of a protein under positive selection. It estimates the ratio of non-synonymous to synonymous substitution rates (K_A/K_S) between a pair of protein-coding DNA sequences, by sliding a 3D window, or sphere, across one reference structure. The program displays the results on the 3D protein structure. In addition, for comparison or when a reference structure is unavailable, the server can also perform a sliding window analysis on the primary sequence. The SWAKK web server is available at http://oxytricha.princeton.edu/SWAKK/.

INTRODUCTION

Mutations and substitutions are fundamental changes in nucleotide sequence over evolutionary time (1). Among the well-established methods for studying the evolution of protein-coding genes, the ratio of non-synonymous substitution rate (K_A, amino acid replacing) to synonymous substitution rate (K_S, silent) is the most powerful measure of selective pressure on a protein (2–8). Since non-synonymous and synonymous substitution sites are interspersed within a gene segment, this approach literally compares the amino acid replacement rate against the silent substitution rate. Traditionally, if $K_A/K_S < 1$, the gene is inferred to be under negative (purifying) selection; if $K_A/K_S = 1$, the gene is probably neutrally evolving; if $K_A/K_S > 1$, the gene is probably under positive (adaptive) selection, since mutations in the gene have higher probabilities of being fixed in the population than expected from the predictions of neutrality.

However, this approach, in effect, averages substitution rates over all amino acid sites in the sequence. Because most amino acids are expected to be under purifying selection, with positive selection most likely affecting only a few sites, this approach often loses the power to detect positive selection. To increase its sensitivity, a sliding window analysis along the primary sequence was introduced (9,10). Recent studies further indicate that when a three-dimensional (3D) protein structure is available, one can detect positive selection much more sensitively by using windows in 3D space instead (11–13). For example, Hughes and Nei (14) detected positive selection at the antigen recognition sites (ARS) in major histocompatibility complex (MHC) alleles but not the whole gene. These sites are close in tertiary space but discontinuous in the primary sequence.

We developed a bioinformatic web server (SWAKK) whose primary purpose is to detect regions under positive selection using a sliding window K_A/K_S analysis (Figure 1). With the input of two protein-coding DNA sequences, one reference protein 3D structure and other user-defined parameters, the web server will automatically align the sequences, calculate K_A/K_S in each 3D window, and display the results on the 3D structure. The server also can perform the analysis on the primary sequence, either for comparison or when a structure is unavailable. In addition, if two inferred ancestral gene sequences are used as an input, the server can examine natural selection in an ancestral branch of a phylogenetic tree (15). We note that two important features distinguish our SWAKK server from other available web servers (16–18) that can identify functionally important sites in proteins. The first difference is that these other web servers focus on each single amino acid site or codon in the multiple sequence alignment, which essentially averages the overall time interval. Instead, our server considers a group of codons within a small window for each pairwise comparison. Second, unlike other web servers where protein 3D structures are only used to display the results, our SWAKK server takes full advantage of the information intrinsically stored in a 3D structure to define neighboring codon groups. Without requiring an explicit evolutionary model or expensive computation, SWAKK thus provides a useful tool to complement the existing arsenal of methods for detecting positive selection.

*To whom correspondence should be addressed. Tel: +1 609 258 1947; Fax: +1 609 258 7892; Email: lfl@princeton.edu

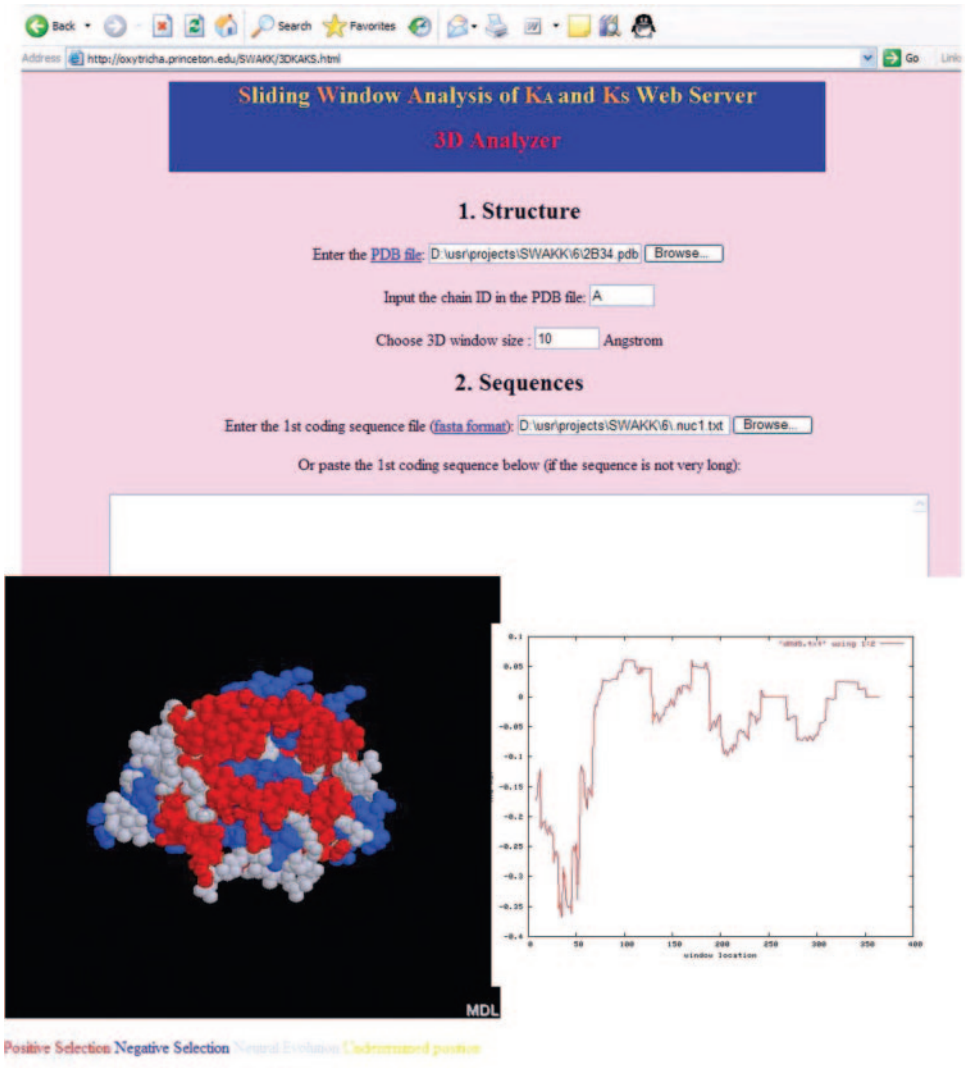

Figure 1. A snapshot of the SWAKK web server and sample output files. The upper part is a snapshot of the 3D analyzer web page. On the bottom are sample output files: Left, 3D provided by the 3D analyzer (when the structure is available), with amino acids colored based on inferred level of selection. Right, 2D graph ($[K_A - K_S]$ versus window location) provided by the 1D analyzer. The example shown is the MHC glycoprotein gene (14,21) and, consistent with the previous studies, most of the sites identified under positive selection are clustered in the ARS domain.

METHODS

SWAKK accepts input as a pair of coding DNA sequences and a reference protein structure (PDB file). The DNA sequences are translated into amino acids and aligned with the amino acid sequence parsed from the PDB file using ClustalW (19). The alignment is then reverse translated to obtain a codon-based sequence alignment. Different translation tables are available to account for variation in genetic codes. Each amino acid in the reference structure is represented as a C^{α} atom. SWAKK constructs 3D windows by placing each amino acid at the center and including all amino acids within a pre-specified distance (in Ångströms) from the center. All the corresponding codons within a window are extracted to form a sub-alignment, and the K_A/K_S score (also the standard error) is calculated using the PAML package (20). Finally, according to the K_A/K_S scores and a user-defined cut-off, the sites (regions) can be classified as positive, negative or neutral, and these are displayed in different colors on the 3D structure using the Chime plug-in component. If a reference structure is not available, the server can also perform the analysis on the primary sequence. In this situation, the window size is defined as the distance in 1D sequence rather than in 3D space, and the results are displayed in the graph drawn by the GNUPLOT software. More detailed information is provided under the links 'Overview', 'Help' and 'FAQ' on the website.

SUMMARY

With more and more protein structures available, we expect this web server to become a valuable bioinformatic tool for detecting functionally important sites. The server facilitates the identification of regions of a protein sequence or structure that may be under positive selection and is easily accessible to the broad biological community.

ACKNOWLEDGEMENTS

The authors thank Dr Yi Zhou for technical assistance and Georgii Bazykin and Landweber lab members for helpful discussion and testing. This work was supported by National Institute of General Medical Sciences grant GM59708 to L.F.L. Funding to pay the Open Access publication charges for this article was provided by NIGMS.

Conflict of interest statement. None declared.

REFERENCES

1. Graur,D. and Li,W.H. (2000) *Fundamentals of Molecular Evolution*. 2nd edn. Sinauer Press, Sunderland, MA.
2. Li,W.H., Wu,C.I. and Luo,C.C. (1985) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol. Biol. Evol.*, **2**, 150–174.
3. Nei,M. and Gojobori,T. (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.*, **3**, 418–426.
4. Li,W.H. (1993) Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J. Mol. Evol.*, **36**, 96–99.
5. Pamilo,P. and Bianchi,N.O. (1993) Evolution of the Zfx and Zfy genes: rates and interdependence between the genes. *Mol. Biol. Evol.*, **10**, 271–281.
6. Comeron,J.M. (1995) A method for estimating the numbers of synonymous and nonsynonymous substitutions per site. *J. Mol. Evol.*, **41**, 1152–1159.
7. Yang,Z. and Nielsen,R. (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.*, **17**, 32–43.
8. Nekrutenko,A., Makova,K.D. and Li,W.H. (2002) The *K*(A)/*K*(S) ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study. *Genome Res.*, **12**, 198–202.
9. Fares,M.A., Elena,S.F., Ortiz,J., Moya,A. and Barrio,E. (2002) A sliding window-based method to detect selective constraints in protein-coding genes and its application to RNA viruses. *J. Mol. Evol.*, **55**, 509–521.
10. Endo,T., Ikeo,K. and Gojobori,T. (1996) Large-scale search for genes on which positive selection may operate. *Mol. Biol. Evol.*, **13**, 685–690.
11. Yang,Z. (2002) Inference of selection from multiple species alignments. *Curr. Opin. Genet. Dev.*, **12**, 688–694.
12. Berglund,A.C., Wallner,B., Elofsson,A. and Liberles,D.A. (2005) Tertiary windowing to detect positive diversifying selection. *J. Mol. Evol.*, **60**, 499–504.
13. Suzuki,Y. (2004) Three-dimensional window analysis for detecting positive selection at structural regions of proteins. *Mol. Biol. Evol.*, **21**, 2352–2359.
14. Hughes,A.L. and Nei,M. (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature*, **335**, 167–170.
15. Zhang,J., Rosenberg,H.F. and Nei,M. (1998) Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proc. Natl Acad. Sci. USA*, **95**, 3708–3713.
16. Lichtarge,O., Bourne,H.R. and Cohen,F.E. (1996) An evolutionary trace method defines binding surfaces common to protein families. *J Mol Biol.*, **257**, 342–358.
17. Glaser,F., Pupko,T., Paz,I., Bell,R.E., Bechor-Shental,D., Martz,E. and Ben-Tal,N. (2003) ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics*, **19**, 163–164.
18. Doron-Faigenboim,A., Stern,A., Mayrose,I., Bacharach,E. and Pupko,T. (2005) Selecton: a server for detecting evolutionary forces at a single amino-acid site. *Bioinformatics*, **21**, 2101–2103.
19. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
20. Yang,Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci.*, **13**, 555–556.
21. Yang,Z. and Swanson,W.J. (2002) Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. *Mol. Biol. Evol.*, **19**, 49–57.